Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments

Gareth J. F. Jones, Ke Zhang and Adenike M. Lam-Adesina

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
email: {gjones,kzhang,adenike}@computing.dcu.ie

Abstract. The Dublin City University participation in the CLEF 2006 CL-SR task concentrated on exploring the combination of the multiple fields associated with the documents. This was based on use of the extended BM25F field combination model originally developed for multi-field text documents. Additionally, we again conducted runs with our existing information retrieval methods based on the Okapi model. This latter method required an approach to determining approximate sentence boundaries within the free-flowing automatic transcription provided, to enable us to use our summary-based pseudo relevance feedback (PRF). Experiments were conducted for the English document collection with topics translated into English using Systran V3.0 machine translation.

1 Introduction

The Dublin City University participation in the CLEF 2006 CL-SR task concentrated on exploring the combination of the multiple fields associated with the speech documents. It is not immediately clear how best to combine the diverse fields of this document set most effectively in ad hoc information retrieval (IR), such as the CLEF 2006 CL-SR task. Our study is based on using the document field combination extended version of BM25 termed BM25F introduced in [1]. In addition, we carried out runs using our existing IR methods based on the Okapi model with summary-based pseudo-relevance feedback (PRF) [2]. Our official submissions included both English monolingual and French bilingual tasks using automatic only and combined automatic and manual fields. Topics were translated into English using the Systran V3.0 machine translation system. The resulting translated English topics were applied to the English document collection.

The remainder of this paper is structured as follows: Section 2 summarises the motivation and implementation of the BM25F retrieval model, Section 3 overviews our basic retrieval system and describes our sentence boundary creation technique, Section 4 presents the results of our experimental investigations, and Section 5 concludes the paper with a discussion of our results.

2 Field Combination

The "documents" of the speech collection are based on sections of extended interviews which are segmented into topically related sections. The spoken documents are provided with a rich set of data fields, full details of these are given in [3]. In summary the fields comprise:

- a transcription of the spoken content of the document generated using an automatic speech recognition (ASR) system,
- two assigned sets of keywords generated automatically (AKW1,AKW2),
- one assigned set of manually generated keywords (MKW1),
- a short three sentence manually written summary of each segment,
- a list of the names of all individuals appearing in the segment.

Two standard methods of combining multiple document fields in retrieval are:

- to simply merge all the fields into a single document representation and apply standard single document field information retrieval methods,
- to index the fields separately, perform individual retrieval runs for each field and then to merge the resulting ranked lists by summing in a process of data fusion.

The topic of field combination for this type of task with ranked information retrieval schemes is explored in [1]. That paper demonstrated the weaknesses of the simple standard combination methods and proposed an extended version of the standard BM25 term weighting scheme referred to as BM25F, which combines multiple fields in a more well founded way.

The BM25F combination approach uses a simple weighted summation of the multiple fields of the documents to form a single field for each document in the usual way. The importance of each document field for retrieval can be determined empirically in separate runs for each field, the count of each term appearing in each field is multiplied by a scalar constant representing the importance of this field, and the components of all fields are then summed to form the overall single field document representation for indexing. Once the fields have been combined in a weighted sum, standard single field IR methods can be applied.

3 System Setup

The basis of our experimental system is the City University research distribution version of the Okapi system [4]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffixes are stripped using the Okapi implementation of Porter stemming [5] and terms are indexed using a small standard set of synonyms. None of these procedures were adapted for the CLEF 2006 CL-SR test collection.

Our experiments augmented the standard Okapi retrieval system with two variations of PRF based on extensions of the Robertson selection value (rsv) for expansion term selection. One method is a novel field-based PRF which we are currently developing [6], and the other a summary-based method [7] used extensively in our earlier CLEF submissions.

3.1 Term Weighting

Document terms were weighted using the Okapi BM25 weighting scheme developed in [4]. With collection frequency weight $cfw(i) = log(((rload + 0.5)(N - n(i) - bigrload + rload + 0.5))/((n(i) - rload + 0.5)(bigrload - rload + 0.5)))$, and $rload = 4$ and $bigroad = 5$. The BM25 k_1 and b values used for our submitted runs were tuned using the CLEF 2005 training and test topics.

3.2 Pseudo-Relevance Feedback

The main challenge for query expansion is the selection of appropriate terms from the assumed relevant documents. For the CL-SR task our query expansion method operates as follows.

Field-Based PRF Query expansion based on the standard Okapi relevance feedback model makes no use of the field structure of multi-field documents. We are currently exploring possible methods of making use of field structure to improve the quality of expansion term selection. For this current investigation we adopted the following method.

The fields are merged as described in the previous section and documents retrieved using the initial query. The Robertson selection value (rsv) [4] is then calculated separately for each field of the original document, but where the document position in the ranked retrieval list has been determined using the combined document. The rsv values in the ranked lists for each field are then normalised with respect to the highest scoring term in each list, and then summed to form a single merged rsv list from the expansion terms are selected. The objective of this process is to favour the selection of expansion terms which are ranked highly by multiple fields, rather than those which may obtain a high rsv value based on their association with a minority of the fields.

Summary-Based PRF The method used here is based on our work originally described in [7], and modified for the CLEF 2005 CL-SR task [2].[1] A summary is made of the ASR transcription of each of the top ranked documents, which are assumed to be relevant for each PRF. Each document summary is then expanded to include all terms in the other metadata fields used in this document index. All non-stopwords in these augmented summaries are then ranked using a slightly modified version of the rsv [4]. In our modified version of rsv(i), potential expansion terms are selected from the augmented summaries of the top ranked documents, but ranked using statistics from a larger number of assumed relevant ranked documents from the initial run.

[1] We refer to this as "Summary-Based PRF" for historical reasons, the "summaries" used here are unrelated to those provided with the CL-SR test collection.

Sentence Selection The summary-based PRF method operates by selecting topic expansion terms from document summaries. However, since the ASR transcriptions of the conversational speech documents do not contain punctuation, we developed a method of selecting significant document segments to identify documents "summaries". This uses a method derived from Luhn's word cluster hypothesis. Luhn's hypothesis states that significant words separated by not more than 5 non-significant words are likely to be strongly related. Clusters of these strongly related significant word were identified in the running document transcription by searching for word groups separated by not more than 5 insignificant words. Words appearing between clusters are not included in clusters, but can be ignored for the purposes of query expansion since they are by definition stop words. The clusters were then awarded a significance score based on two measures:

- Luhn's Keyword Cluster Method: Luhn's method assigns a significance score to each word cluster.
- Query-Bias Method: Assigns a score to each word cluster based on the number of query terms in each one.

The overall score for each word cluster was then formed by summing these two measures for each one. The word clusters were then ranked by score with the highest scoring ones selected as the segment summary.

4 Experimental Investigation

This section gives results of our experimental investigations for the CLEF 2006 CL-SR task. We first present results for our field combination experiments and then those for experiments using our summary-based PRF method. Results are shown for precision at rank cutoff 5, 10 and 30 documents, standard mean average precision (MAP) and the total number of relevant documents retrieved summed across all topics down to rank of 1000 documents.

For our formal submitted runs the system parameters were selected by optimising results for the CLEF 2005 CL-SR training and test collection. Our submitted runs for the CLEF 2006 are indicated by a * in the tables.

4.1 Field Combination Experiments

Two sets of experiments were carried out using the field combination method. The first uses all the document fields combining manual and automatically generated fields, and the other only the automatically generated fields. We report results for our formal submitted runs using our field-based PRF method and also baseline results without feedback. We also give further results obtained by optimising performance for our systems using the CLEF 2006 CL-SR test set.

Table 1. All fields results with parameters set using CLEF 2005 data.

	TD	Recall	MAP	P5	P10	P30
Monolingual	Baseline:	1844	0.223	0.366	0.293	0.255
English	PRF*	1864	0.202	0.321	0.288	0.252
Bilingual	Baseline	1491	0.158	0.306	0.256	0.204
French-English	PRF*	1567	0.160	0.291	0.252	0.199

Table 2. All fields results with parameters optimised using CLEF 2006 topics.

	TD	Recall	MAP	P5	P10	P30
Monolingual	Baseline:	1908	0.234	0.364	0.342	0.303
English	PRF	1929	0.243	0.364	0.370	0.305
Bilingual	Baseline	1560	0.172	0.315	0.267	0.231
French-English	PRF	1601	0.173	0.315	0.267	0.225

Table 3. Auto only fields results with parameters set using CLEF 2005 data.

	TD	Recall	MAP	P5	P10	P30
Monolingual	Baseline	1290	0.071	0.163	0.163	0.149
English	PRF*	1361	0.073	0.152	0.142	0.146
Bilingual	Baseline	1070	0.047	0.119	0.113	0.106
French-English	PRF*	1097	0.047	0.106	0.094	0.102

Table 4. Auto only fields results with parameters optimised using CLEF 2006 topics.

	TD	Recall	MAP	P5	P10	P30
Monolingual	Baseline	1335	0.080	0.224	0.215	0.169
English	PRF	1379	0.094	0.188	0.206	0.184
Bilingual	Baseline	1110	0.050	0.121	0.127	0.123
French-English	PRF	1167	0.055	0.127	0.142	0.124

All Field Experiments

Submitted Runs Based on development runs with the CLEF 2005 data the Okapi parameters were set empirically as follows: $k_1 = 6.2$ and $b = 0.4$, and document fields were weighted as follows: Name field × 1; Manualkeyword field × 10; Summary field × 10; ASR2006B × 2; Autokeyword1 × 1; Autokeyword2 × 1.

The contents of each field were multiplied by the appropriate factor and summed to form the single field document for indexing. Note the unusually high value of k_1 arises due to the change in the $tf(i, j)$ profile resulting from the summation of the document fields [1].

Results for English monolingual and bilingual French-English runs are shown in Table 1. For both topic sets the top 20 ranked terms were added to the topic for the PRF run with the original topic terms upweighted by 3.0.

It can be seen from these results that, as is usually the case for cross-language IR, performance for monolingual English is better than bilingual French-English for all measures. A little more surprising is that while the application of the

field-based PRF gives a small improvement in the number of relevant documents retrieved, there is little effect on MAP, and precision at high ranked cut off points is generally degraded. PRF methods for this task are the subject of ongoing research, and we will be exploring these results further.

Further Runs Subsequent to the release of the relevance set for the CLEF 2006 topic set further experiments were conducted to explore the potential for improvement in retrieval performance when the system parameters are optimised for the topic set. We next show our best results achieved so far using the field combination method. For these runs the fields were weighted as follows: Name field × 1; Manualkeyword field × 5; Summary field × 5; ASR2006B × 1; Autokeyword1 × 1; Autokeyword2 × 1, and the Okapi parameters set empirically as follows: $k_1 = 10.5$ and $b = 0.35$.

The results for English monolingual and bilingual French-English runs are shown in Table 2. For monolingual English the top 20 terms were added to the topic for PRF run with the original topic terms upweighted by 33.0 For the French bilingual runs the top 60 terms were added to the topic with the original terms upweighted by 20.0. For these additional runs all test topics were included in all cases.

Looking at these additional results it can be seen that parameter optimisation gives a good improvement in all measures. It is not immediately clear whether this arises due to the instability of the parameters of our system, or a difference in some feature of the topics between CLEF 2005 and CLEF 2006. We will be investigating this issue further. Performance between monolingual English and bilingual French-English is similar to that observed for the submitted runs. PRF is generally more effective or neutral with the revised parameters, again we plan to conduct further exploration of PRF for multi-field documents to better understand these results.

Automatic Only Field Experiments

Submitted Runs Based on development runs with the CLEF 2005 CL-SR data the system parameters for the submitted automatic only field experiments were set empirically as follows: $k_1 = 5.2$ and $b = 0.2$ and the document fields were weighted as follows: ASR2006B × 2; Autokeyword1 × 1; Autokeyword2 × 1.

These were again summed to form the single field document for indexing. The same French-English topic translations were used for the automatic only field experiments as for the all field experiments.

The results of English monolingual and bilingual French-English runs are shown in Table 3. The top 30 terms were added to the topic for PRF run with the original topic terms upweighted by 3.0. From these results it can again be seen that there is the expected reduction in performance between monolingual English and bilingual French-English. The field-based PRF is once again shown generally not to be effective with this dataset. There is a small improvement in the number of relevant documents retrieved, but no there is little positive impact for precision.

Table 5. Results for monolingual English with all document fields.

		Recall	MAP	P10	P30
TDN	Baseline	1871	0.279	0.449	0.367
	PRF†	1867	0.291	0.461	0.377
TD	Baseline	1819	0.241	0.436	0.321
	PRF	1885	0.259	0.415	0.348

Table 6. Results for monolingual English with auto document fields.

		Recall	MAP	P10	P30
TDN	Baseline	1259	0.070	0.182	0.160
	PRF	1252	0.072	0.194	0.155
TD	Baseline	1247	0.058	0.124	0.122
	PRF	1255	0.066	0.146	0.129

Further Runs Subsequent to the release of the relevance set for the CLEF 2006 topic set further experiments were again conducted with system system parameters optimised using the test data set. For these runs the field weights were identical to those above for the submitted runs, and the Okapi parameters were modified as follows: $k_1 = 40.0$ and $b = 0.3$.

The results of English monolingual and bilingual French-English runs are shown in Table 4. For monolingual English the top 40 terms were added to the topic for PRF run with the original topic terms upweighted by 3.0 For the bilingual French-English runs the top 60 terms were added to the topic with the original terms upweighted by 3.5.

Once again optimising the system parameters results in an improvement effectiveness of the PRF method. Further experiments are planned to explore these results further.

4.2 Summary-Based PRF Experiments

For these experiments the document fields were combined into a single field for indexing without application of the field weighting method. The Okapi parameters were again selected using the CLEF 2005 CL-SR training and test collections. The values were set as follows: k_1=1.4 and b=0.6. For all our PRF runs, 3 documents were assumed relevant for term selection and document summaries comprised the best scoring 6 word clusters. The rsv values to rank the potential expansion terms were estimated based on the top 20 ranked assumed relevant documents. The top 40 ranked expansion terms taken from the clusters were added to the original query in each case. Based on results from our previous experiments in CLEF, the original topic terms are up-weighted by a factor of 3.5 relative to terms introduced by PRF.

All Field Experiments Table 5 shows results for monolingual English retrieval based on all document fields using TDN and TD field topics.[2] Rather surprisingly the baseline result for this system is better than either of those using the field-weighted method shown in Table 1. The summary-based PRF is shown to be effective here, as it has in many previous submissions to CLEF tracks in previous years. TDN topics produce better results than the TD only topics.

Automatic Only Field Experiments Table 6 shows results for monolingual English retrieval based on only auto document fields again using TDN and TD topic fields. The summary-based PRF method is again effective for this document index. However, the results are not as good as those in Table 3 using the field weighted document combination method.

5 Conclusions and Further Work

This paper has described results for our participation in the CLEF 2006 CL-SR track. We explored use of a field combination method for multi-field documents, and two methods of PRF. Results indicate that further exploration is required of the field combination approach and our new field-based PRF method. Our existing summary-based PRF method is shown to be effective for this task. Further work will also concentrate on combining effective elements of both methods.

References

[1] Robertson, S. E., Zaragoza, H., and Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields, Proceedings of the 13th ACM International Conference on Information and Knowledge Management, pages 42-49, 2004.

[2] Lam-Adesina, A. M., and Jones, G. J. F.: Dublin City University at CLEF 2005: Cross-Language Speech Retrieval (CL-SR) Experiments, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.

[3] White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang, X.: Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.

[4] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu,M. M. and Gatford, M.:Okapi at TREC-3, Proceedings of the Third Text REtrieval Conference (TREC-3), pages 109-126. NIST, 1995.

[5] Porter, M. F.: An Algorithm for Suffix Stripping, Program, 14:10-137, 1980.

[6] Zhang, K.: Cross-Language Spoken Documenr Retrieval from Oral History Archives, MSc Dissertation, School of Computing, Dublin City University, 2006.

[7] Lam-Adesina, A. M., and Jones, G. J. F.: Applying Summarization Techniques for Term Selection in Relevance Feedback, Proceedings of the Twenty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-9, New Orleans, 2001. ACM.

[2] Our submitted run using this method was found to be in error, and the corrected one is denoted by the † in this table.